**Marketing Analytics : Recommender Systems Project**

*Kimberly-Clark is planning to promote the Huggies brand targeting customers who are currently buying another brand.*

28th February 2019

Capstone Group 2 : Kwaku Danso-Manu, Eric Gu, Alex Lu, Alex Su



Table of Contents

1. Background and Introduction
2. Pre-processing Data
3. Data Exploration
4. Exploring Customers
5. Market Level
6. Huggies vs Not Huggies
7. All Customers
8. Competitors
9. Huggies
10. Recommendations with Collaborative Filtering
11. Considering Time Factor in Analysis
12. Expected Revenue
13. Rankings
14. Deployment

**Background and Introduction**

While Kimberly-Clark is a massive organization with several product lines, they are currently losing ground on the baby products market. Our goal for this project is to find customers who are currently buying other brands of products and find ways to get them buying Huggies line of products.

**Pre-Processing Data**

The data provided is 5GB worth of information from Pernalonga stores which contains detailed information on products, stores and customers. As from our previous analysis, the transaction IDs did not suffice and we had to create a new unique transaction Id for each observation. This resulted in 2,9617,585 unique transactions with 10,770 unique products and 7920 customers.

We then limited the scope of the data by focusing only on baby related products as our main goal is to promote huggies brands of products to other customers. The product categories include diapers, baby food, baby hygiene, baby shampoos, baby conditioner and baby bath products.

We then filter the data to focus on the baby product market as a whole, huggies branded products and competitor products for exploration and further understanding of the customers.

**Data Exploration**

In the baby market, there are 14 brands in total sharing the market including Huggies. In terms of quantity for product sold, Huggies’ market share is 1.9 %.

|  |  |  |
| --- | --- | --- |
|  | Huggies | Market |
| Quantity of Product Sold | 2860 | 152158 |
| Customer Number | 1205 | 7452 |

To take a closer look into the market which includes only the categories Huggies is currently selling which are “OUT. BABY HYGIENE ITEMS” and “DIAPERS”, there are 7 brands including Huggies in this more focus market. Huggies’s market share is 2.1% which is relatively small comparing to other brands in the market.

By differentiating the market into “OUT. BABY HYGIENE ITEMS” and “DIAPERS”, we are able to explore even deeper in the specific type of product to understand the situation of Huggies in the market. From the below table, the average unit price in terms of “OUT. BABY HYGIENE ITEMS” product is higher than the average market price. On the other hand, the average unit price in “DIAPERS” product is slightly lower than the average market price.

|  |  |  |
| --- | --- | --- |
|  | Huggies | Product Focused Market |
| Price per Unit  (OUT. BABY HYGIENE ITEMS) | $4.15 | $1.70 |
| Price per Unit  (DIAPERS) | $8.53 | $9.41 |

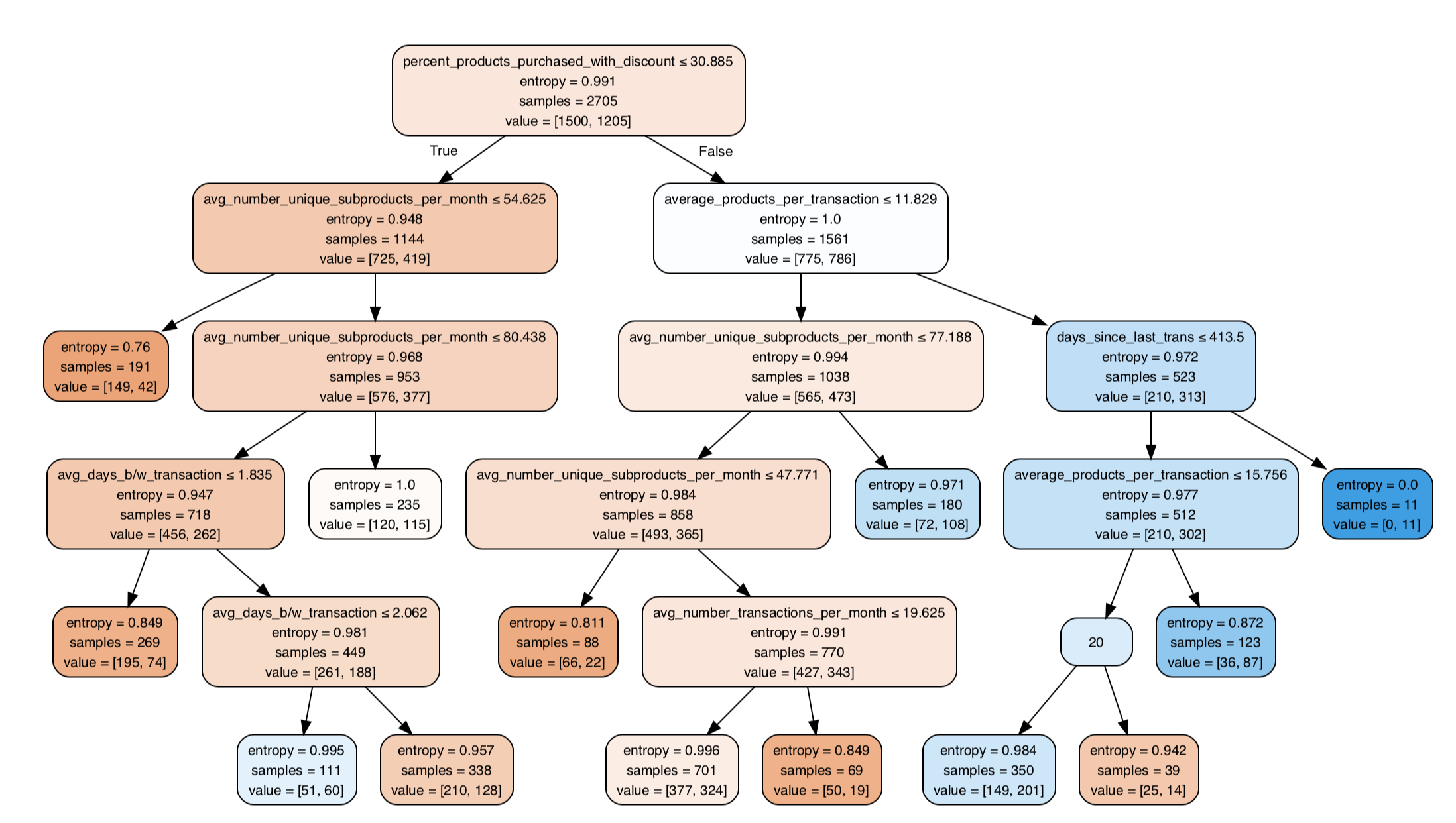
**Exploring Customers**

**Market Level : Buy Huggies vs Other Brands**

Next, we decided to explore the characteristics of customers in the baby products market. We conducted analysis on several levels and will delve further below. The main goal of this process was to examine and come up with business rules to help with understanding our customers and ways to help convert customers who are currently not buying Huggies branded products.

First, we looked at the baby products market level, and we examined customers who buy huggies branded products versus those who don’t. Our exploratory analysis revealed some interesting insights. There are two clear groups on each side, with the first and most important deciding factor being the the presence of discounts on a product. Huggies being considered a premium brand, is purchased without discounts as these customers trust its quality and durability. Huggies buyers also tend to do more general shopping as they get these products and visit the stores more often. The competitor pool seem to have the profile of customers who prefer to buy their products usually only when they are discounted. They also seldom buy several different products which ties in with the notion that they target when these baby products are discounted and buy those only or with a few other items.

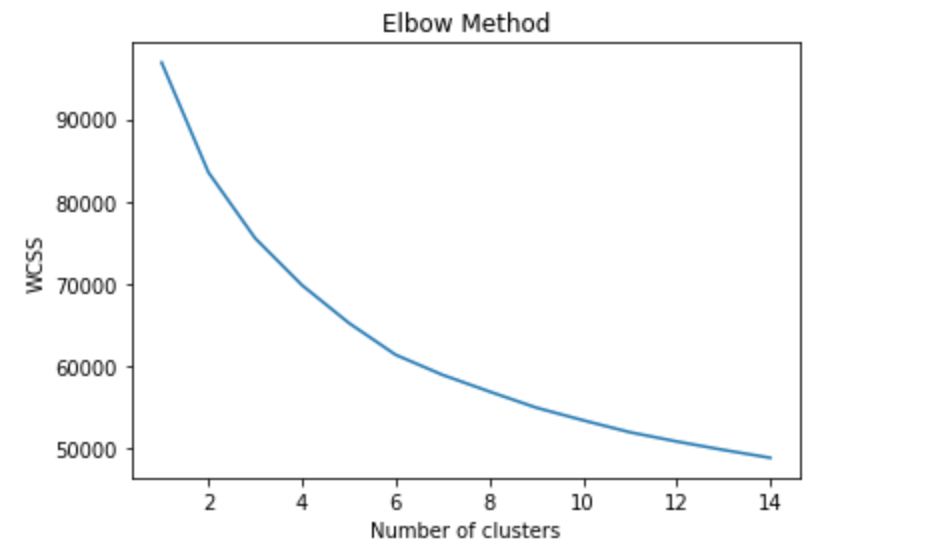
The summarized results are shown in the figure below.



**Market Level : General Customer Segments**

Next, we analyzed the general profiles of customers in the whole market to compare how their characteristics match with the profile of huggies customers discovered above.

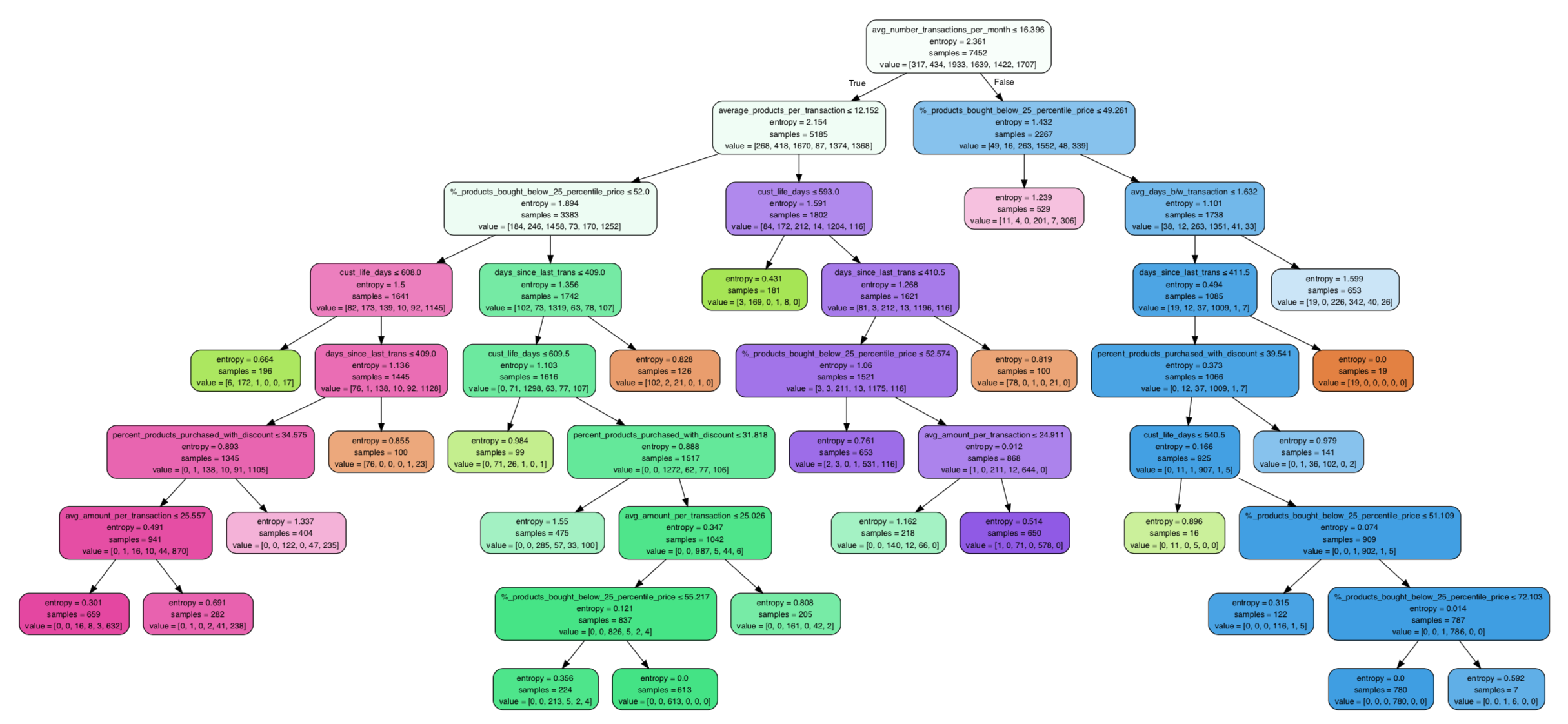
We first conducted an analysis on the data to find out the optimal number of groups to divide the customer pool into. We used the K-Means clustering algorithm which compares similarity of customers based on the attributes we provided. It also gives us a suggested number of groups for our analysis with a graph known as the elbow plot, seen below, with the sharp edge point indicating the number of groups.



The plot indicates that we should group our customers into 6 segments. Our first inclination was that this could be the division of the six categories of baby products in the data, discussed above. Five main categories are dominant with one having few instances and tapering out early on in the tree. The group observed to be closest to huggies customers are colored in blue. They rarely buy products below the 25 percentile price, shop frequently and are not sensitive to discounts. Extracting the customers in this field and finding out the section who are not huggies customers and targeting them would be a great approach as they tend to be very similar to the customers who buy huggies branded products.

The group of customers we should avoid targeting as they are far from our customer profile are colored in pink. Characteristically penny pinchers who rarely go shopping and usually by products on discount.

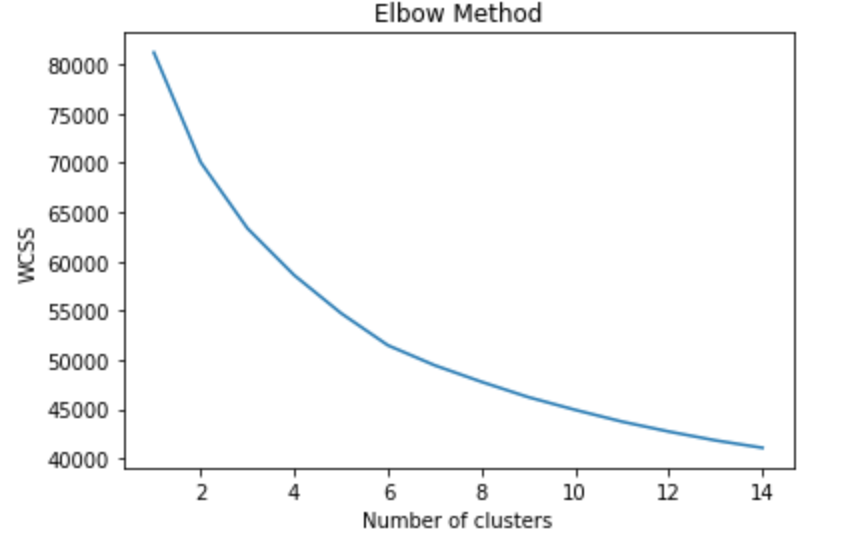
A visual of their characteristics is shown below.



**Competitors Only**

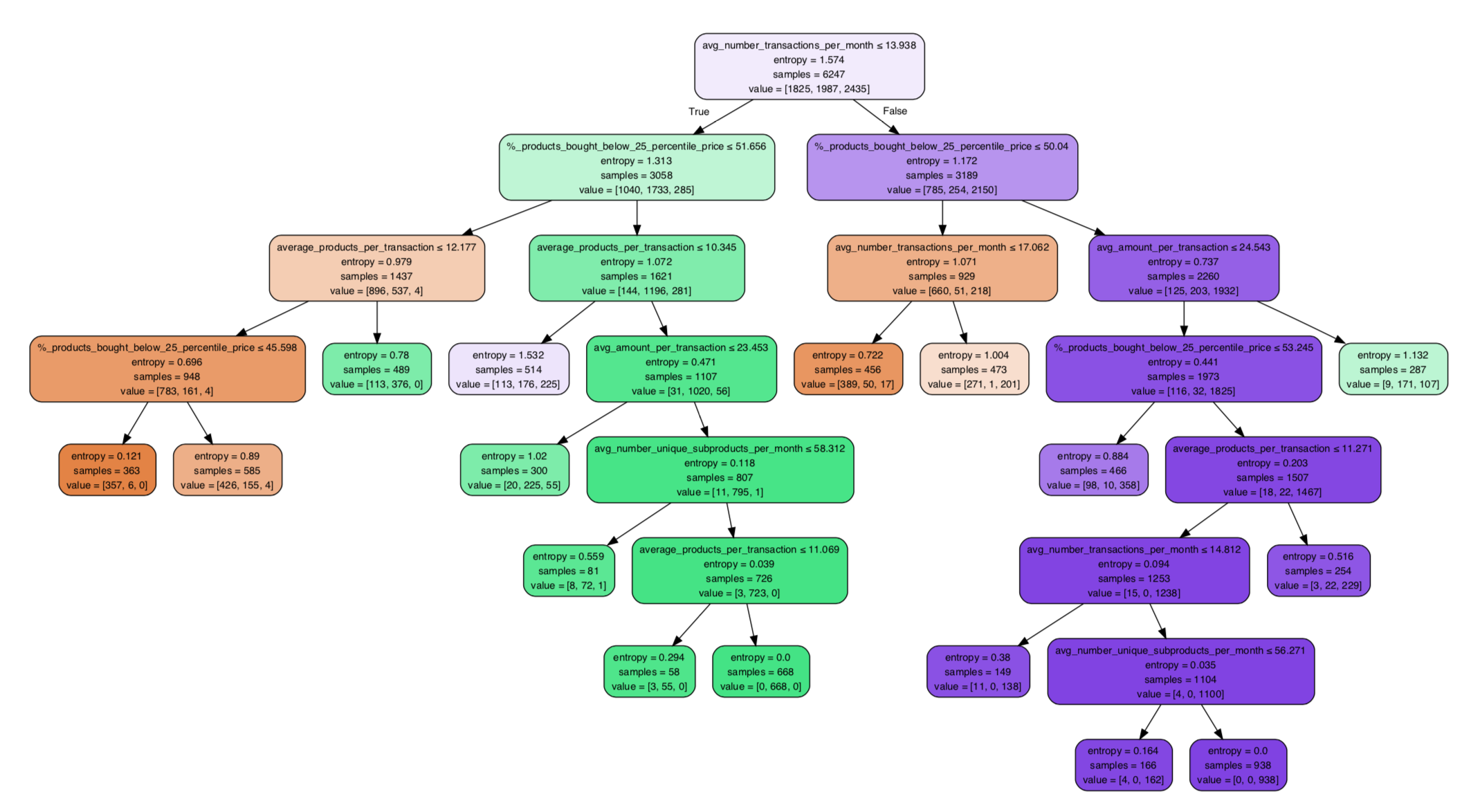
After analyzing the whole market, we get a clearer sense of what our ideal profile looks like for customers we want to target. We now analyze those who are currently not buying Huggies products and see if there is a smaller sect who match our profile.

As before, we let the K-Means algorithm together with the elbow plot indicate how many groups these customers naturally split into and the diagram is displayed below.



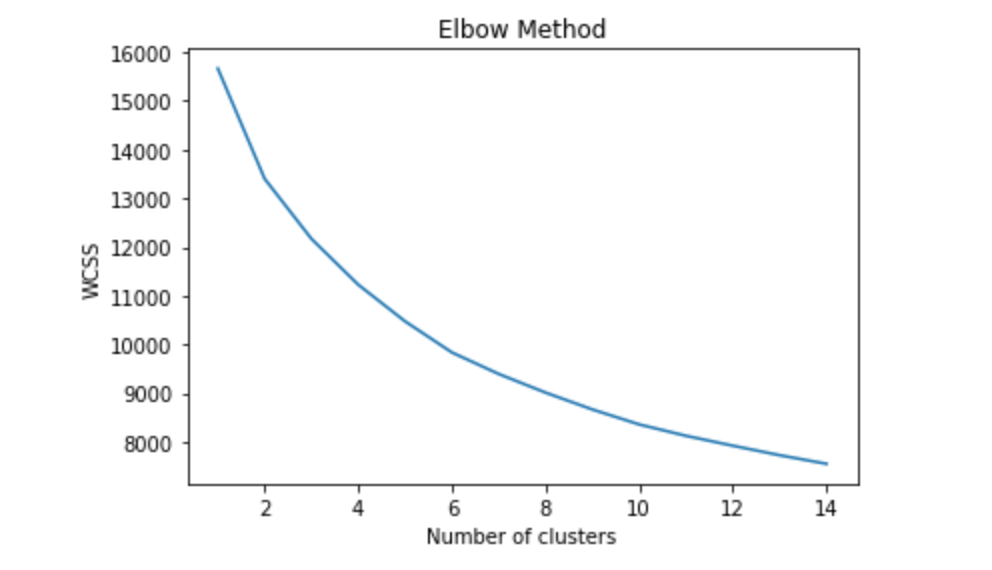
The sharp edge appears to be at 3 clusters and this is the number used for our segmentation. The third group in purple seem to be very similar to our customer profile. A group of people who are also insensitive to discounts, buy several distinct products which each transaction but tend not to spend a lot per transaction. We would have to factor this into consideration when designing our personalized promotion schemes for the customers in question.

The defining features of the three segments are broken down in the decision tree below.

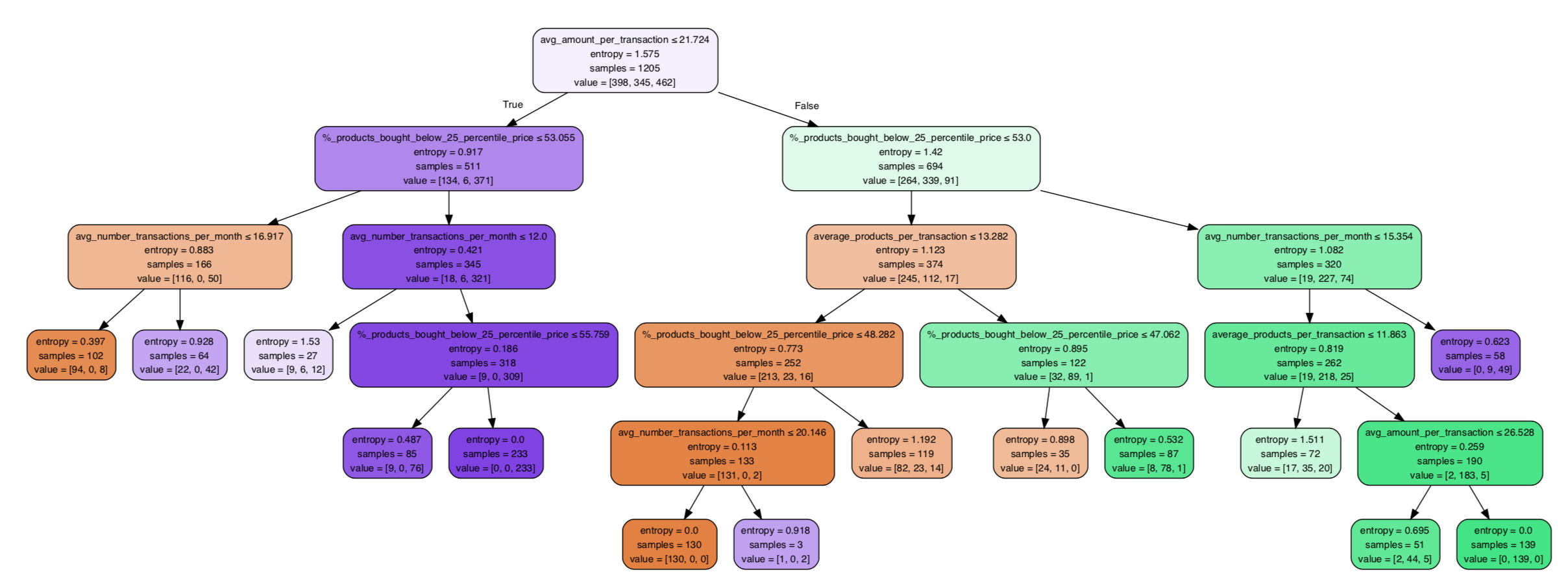


**Huggies Only**

Finally, we decided to validate our customer profile by also examining the various profiles of customers currently buying huggies products. Using the same approach described above, we see the elbow plot of below.

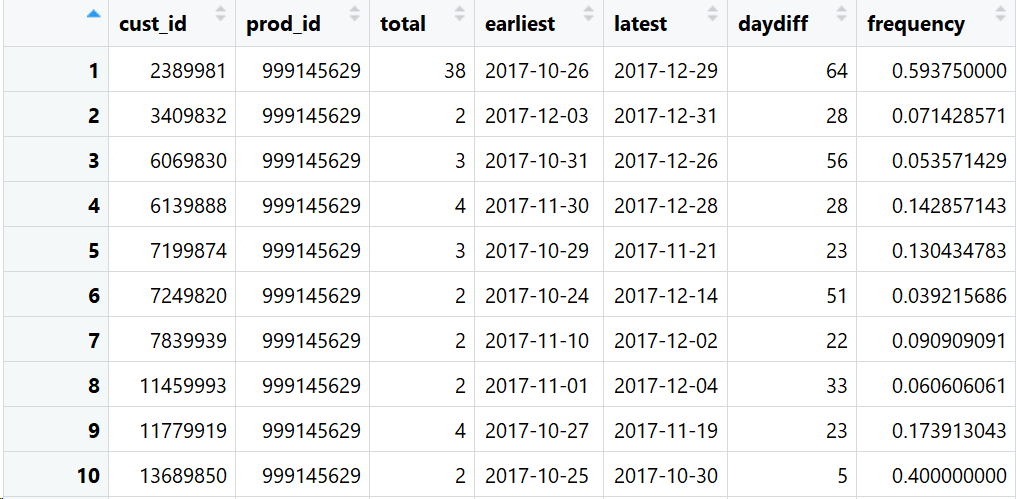


With an incline towards 3 segments, we grouped the customers and analyzed how they differ from each other. At first glance, we notice that the main differentiating factor being the amount spent per transaction. They are all generally frequent shoppers, with those visiting the stores less having a higher amount spent per transaction which is a logical flow.



**Product Recommendations**

There are a total of 7452 customers in the baby products market. For each customer, we calculated the total quantity of each baby product the customer has bought and named it ‘total’. For each product, we also found the earliest and latest purchase date of the customer. The difference in days is then calculated and we named it ‘daydiff’. By dividing the ‘total’ by ‘daydiff’, we were able to find the purchasing frequency of each product for a given customer. The result is shown below where the last column, ‘frequency’, indicates the purchasing frequency in days.



The reason we didn’t calculate frequencies based on weeks or months is that if a customer purchased the same product multiple times within one week or one month, the value of ‘total’ would be a positive number but the difference in week or month would be zero, resulting in a zero in the denominator. We are assuming that each customer would only make one transaction on a particular product in a given day, so we excluded those one-time purchase rows where the earliest transaction date equals the latest. We did this for two reasons:

1. Those one-time purchases are not representative to delineate a customer’s preference because those might be arbitrary purchases and since those customers never returned for another purchase, it somehow indicated they don’t like or they don’t need those products.

2. By excluding those rows, we no longer had any cells with values of zero for ‘daydiff’. After data cleaning, we ended up with 5826 customers and 129 baby products.

Then, we created an affiliation matrix of customer vs product and used purchasing frequencies as the ratings. By using collaborative filtering techniques, we would be able to find the customers who are most likely to buy Huggies products. Customers who have similar purchasing frequencies on certain products might have similar preferences in terms of baby products and their babies might be similar in age.

The reason we limited our product range to only baby products when calculating similarity is that we were concerned of noise. Two customers can be 95% similar in terms of purchasing behavior in all the other products but with completely different preferences when it comes to baby products. Therefore, we only focused on baby products to get unbiased ratings for each customer.

We tried both user-based collaborative filtering and item-based collaborative filtering. And for each type of collaborative filtering, we compared the performance in terms of RMSE, MSE and MAE among using Jaccard similarity, cosine similarity and Pearson similarity. The result suggests that user-based collaborative filtering with Pearson similarity has the best overall performance score.

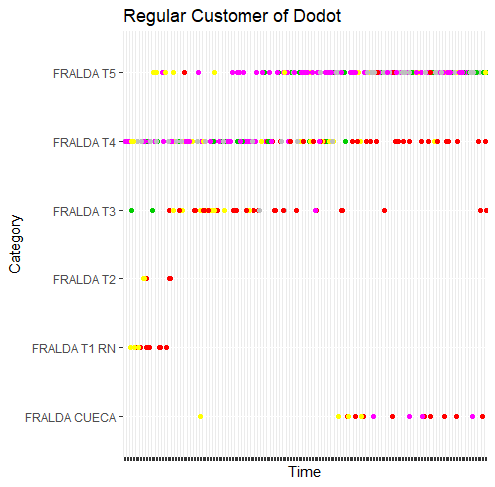
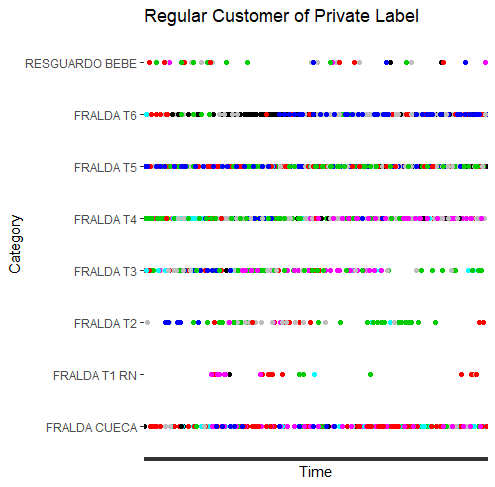
We proceeded with this method and for each Huggies product, the algorithm produced a list of ratings for customers who’s not currently buying the product. A higher rating indicates that a customer is more likely to buy the product.

**Considering Time Factor in Analysis**

For wipers, there’s no age constraint on it because from our research and personal experience, people at different ages, with or without babies also buy baby wipers because of their quality. However, due to the nature of diapers, there is a certain age at which the baby becomes old enough, the customer will stop purchasing them. As a result, we tried to identify which customer should be excluded based on whether their baby is old enough.

In order to do so, we first need to identify which products are the largest size of the their category. We looked for regular customers who bought all the categories of a product at least once within a week. We then plot all the transactions with different color for each customer on a timeline to see if there is a sequence in terms of the purchase made for a specific category of product. We then analyzed the subcategories (T1 ~ T6) for the baby products which indicated the size of baby diaper the product was. With this knowledge, we could conclude that these larger products that were bought with less frequency after frequent purchase of smaller sizes would not be customers we would want to target.

Taking a look at the graphs shown below, with the Dodot brand as an example, with time on the horizontal axis and product categories on the vertical. There are patterns in terms of the sequence for purchasing diaper. For diaper T1, by looking at the customer as the red dot, T1 usually be bought before T2, T3, and T4. Therefore, we can validate that T1 is the smaller size comparing to T2 and T3. By looking at the customer as the purple dot, T4 seems to be bought earlier than T5. So, we determine that T5 is the largest and T1 is the smallest. For the category “cueca”, since the pattern is not clear enough, we treat that as “other” category which cannot be an indicator to determine which customers need to be excluded in 2018. The approach is exactly the same with Private Label diaper which T5 and T6 is the biggest size of diaper, and besides “cueca”, we consider “Resguardo BEBE” to be “other” as well because there is no clear pattern.

After we found the biggest size of the product, we located those customers who bought the biggest size of the diapers for at least 180 days for minimum 6 times before 6/1/2017. We consider them as customers who no longer need diapers in April 2018 because their baby is old enough. We set the condition higher that baby should be using the biggest diaper for at least 14 months so that we can be more confident that the list includes all the customers who still need diaper related products. The result showed us 144 customers whose babies are big enough to be eliminated in the list for recommendation. We filtered out these 144 customers with each of our rating tables to have a final list for the potential customers who are likely to buy Huggies products. For those babies who were not eliminated, we went on to estimate expected values of each of them switching to Huggies.

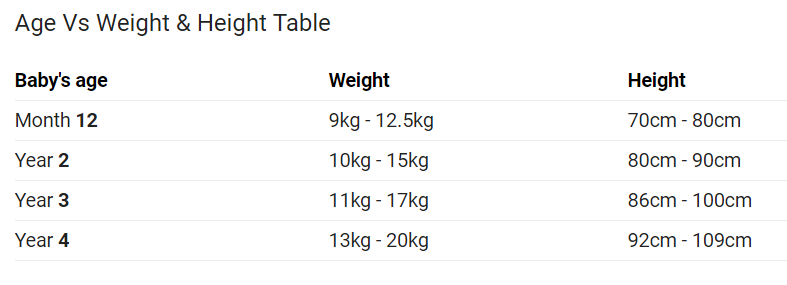
**Expected Revenue**

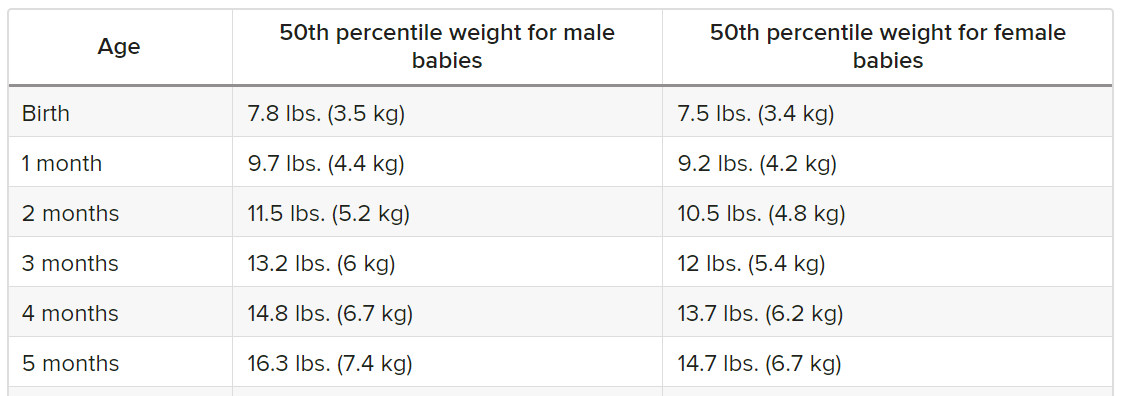
Diapers as a product possesses some interesting features regarding user behavior. If someone has a baby of 0 to 3 years old, they have to buy diapers, and the frequency of purchase do not vary greatly amongst users. Therefore, we incorporated a “expected revenue” factor in our analysis.

For each customer who has bought a diaper at Pernalonga, we estimated their baby’s age based on their buying behavior. We discovered that the product subcategory column (contains a code from T1 - T6) for the diaper category describes which baby ages this diaper product serve.

(“Talla 6” and “Talla 5” at the lower left corner corresponds to subcategories. Baby weights that corresponds to Talla 6 and Talla 5 are indicated in kilograms.)





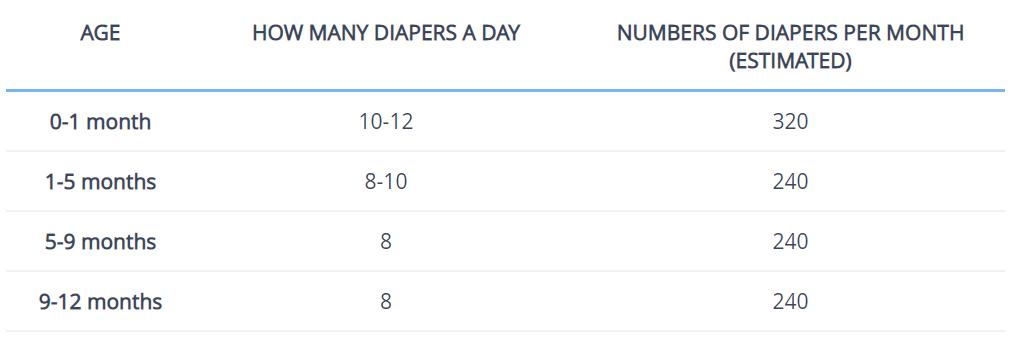


We assumed that when a customer switched from one subcategory to the next (e.g. Talla 5 to Talla 6) is when their baby has reached one of the critical ages indicated in the above chart. For customers who have only bought in one subcategory or have only bought in the two special categories (for diaper underwear and absorbent bed sheets), we estimated their babies’ age at April 2018 with the first time they purchased diapers in Pernalonga.

Using age, we then estimated how many more diapers each baby is expected to use. From online resources including articles and parenting forums, we estimated the numbers of diapers per day babies use at different ages (at the month level because the accuracy of baby age estimates isn’t reliable at the day level).

|  |  |  |  |
| --- | --- | --- | --- |
| Age Range | Diapers/Day | Diapers/Month | Sub-category |
| 0mo – 1mo | 12 | 360 | T1 |
| 2mo - 5mo | 10 | 300 | T1 - T2 |
| 6mo – 12mo | 8 | 240 | T2 |
| 1yr – 2yr | 6 | 180 | T3 |
| 2yr – 3yr | 6 | 180 | T4 |

With the average amounts paid per bag of diapers, the future quantity estimated above, we calculated an expected revenue for each customer who has bought a diaper at Pernalonga. Using the minimum sales prices of products to estimate the product costs, we transformed expected revenues to expected profits.



<https://www.newkidscenter.com/how-many-diapers-a-day.html>

**Ranking**

Remember that our target list of customers include those who have bought baby products but not diapers at Pernalonga. The reasoning being we wanted to identify parents who are buying Huggies’ competitors’ products at other stores. Inevitably, when we target parents who are buying diapers at other stores, we will include people who are already buying huggies. However, according to Huggies’ market share estimation using Pernalonga data, falsely targeted customers can be limited to less than 2% of customers buying diapers from other stores.

In the previous step, we only estimated baby age for customers who are known parents, or in other words, customers who have bought diapers at Pernalonga. Therefore for each Huggies product, we will do two rankings, one for customers who have bought diapers in Pernalonga and will be calculated from a probability of switching times customer expect value, another for customers who we think might be buying diapers in other stores and will only be ranked according the probabilities produced in the recommendations section. A number of assumptions were made when estimating the age for each babies, thus we do not want to eliminate customers from the targeting pool using the age estimates at this step. As a result, customers whose expected values of converting are zero will also be ranked according the probabilities produced in the recommendations section.

**Deployment**

Following this analysis, business decisions will be made regarding how many customers to target depending on how permissive or conservative stakeholders are. These decisions will be transformed into cut-off values in the ranked lists. Lower cut-off points will result in targeting more customers but marketing costs would also rise. Higher cut-off points will target more customers and reduce marketing costs. Kimberly-Clark can do a cost-benefit analysis to determine the cut-offs. If the marginal cost of promoting a certain product is low, they should consider including as many customers as possible on the list during the marketing campaign.

Customers to be targeted will be classified into the three segments determined earlier and targeted with suited strategies.

**Segment 1**: Customers who buy regardless of discounts, shop frequently, and buy diverse portfolios of product in their transactions

Strategy: Emphasize that Huggies is a brand with premium quality; improve awareness.

**Segment 2**: Customers who don’t shop frequently and really care about discounts (cherry-pickers)

Strategy: Kimberly-Clark should offer them coupons and promotions on Huggies products. For example, customers can sign-up on Huggies website and get free shipping on qualified purchases; For every 3 pack of Huggies diapers they buy, they can get 1 pack of Huggies wipers for free,etc.

**Segment 3**: Customers who consider price and quality concurrently and are moderate in their purchase frequencies.

Strategy: Pending further investigation of this customer segment, these customers will likely be targeted with a hybrid strategy that emphasized balance.